9M

УДК 681.3

А.А. Марченко, А.А. Никоненко
Киевский национальный университет имени Тараса Шевченко, г. Киев, Украина
rozenkrans@yandex.ru, andrey.nikonenko@gmail.com

Контекстный семантический анализ текста. Система текстового мониторинга и качественного оценивания фокусного объекта

В данной работе содержится описание системы ассоциативно-семантического контекстного анализа естественноязыковых текстов, на базе которой реализована прикладная система мониторинга текстовых потоков и корпусов с блоком качественного оценивания лингвистических фокусных объектов, предназначенная для вычисления различных качественных характеристик и параметров заданных объектов и процессов.

Контекстный семантический анализ текста

Система ассоциативно-семантического контекстного анализа ЕЯ-текстов разработана на основе глобальной мультилингвистической лексико-семантической онтологической базы знаний UkrRusWordNet, поддерживающей лексиконы украинского, русского и английского языков. Функции и процедуры семантического анализа реализованы на базисе гиперсети данной онтологии. Процесс ассоциативно-семантического анализа в системе можно условно разделить на три этапа:

– переход от слов и словосочетаний предложений до соответствующих семантических значений – концептов онтологии;
– сборка семантических фреймов предложений текста;
– объединение семантических структур предложений текста в единую семантическую сеть текста.

На первом этапе система определяет в семантической сети онтологической базы знаний концепт, соответствующий корректному значению слова или словосочетания в тексте. Эта задача решается поиском того значения слова из множества возможных альтернатив концептов, которое семантически является наиболее близким к значениям слов-соседей из локального окружения данного слова. Степень близости находится поиском кратчайшего пути между концептами в сети онтологии [1], [2]. Таким образом, вычисляется расстояние в онтологии между предполагаемым значением слова и концептами-значениями слов ближайшего окружения. В системе реализована алгоритмическая модель контекстного анализа.

Второй этап – построение семантического фрейма текущего предложения входного текста. Он заключается в заполнении слотов фреймовой структуры предложения. Выбор типа слота для заполнения значением концепта слова зависит от синтаксической позиции этого слова в грамматической структуре предложения. Заполнение слотов выполняется путем анализа дерева разбора предложения и

синтаксических позиций слов и словосочетаний для каждого концепта с использованием семантико-синтаксических таблиц модально-ролевых падежей, подобных падежам Филмора [3].

Третья фаза смыслового анализа – объединение изолированных семантических фреймов предложений в связную семантическую сеть текста. Объединение двух структур в одну сеть выполняется по принципу объединения семантически тождественных вершин, то есть если в структурах $G1$ и $G2$ есть вершины, которые ссылаются на один семантический концепт, эти вершины объединяются в одну.

На выходе системы генерируется семантическая сеть входного текста, которая содержит в вершинах концепты текста, связанные дугами семантических отношений. Дальнейшая смысловая обработка полученной семантической сети текста позволяет решать широкий класс задач компьютерной лингвистики.

Был предложен метод моделирования семантического контекста и вычисления семантической контекстной близости значений слов с использованием онтологической базы знаний. Онтология является основой семантического анализа тем семантическим полем, в рамках которого можно вычислять смысловую близость семантических интерпретаций лексем текста относительно ближайшего окружения, то есть контекста. Это и есть отправная точка для моделирования такого ключевого явления как языковый контекст вообще и контекстного анализа ЕЯ-текстов в частности.

Онтология является иерархической семантической сетью, вершины которой содержат концепты (смысловые единицы), а дуги – это семантические отношения между концептами. Семантика (смысл) концепта описывается его смысловыми отношениями с другими концептами сети. Реляционная позиция концепта в онтологической сети описывает его семантическое значение, свойства, отношения с другими концептами и все другие характеристики, которые возможно передать естественным языком. Онтологические технологии используют лингвистические модели представления знаний об окружающем мире и предметных областях для эффективной записи и обработки информации ЕЯ-типа [4].

Слова и словосочетания языка сохраняются в лексиконе системы. Каждая лексема в системе ссылается на множество значений, которые ей соответствуют в данном языке. Слово, употребленное вне контекста, может иметь любое значение из множества концептов, которые прописаны ему в онтологической базе знаний. Если слово употребляется в контексте определенного предложения, то его значение должно согласовываться со значениями слов, которые стоят рядом. Семантические значения слов предложения должны создавать смысловое единство в структуре семантического фрейма предложения. Поэтому значения концептов слов, которые стоят рядом в предложении, должны быть семантически как можно ближе друг к другу.

На вход блока контекстного анализа подается последовательность слов w1 w2 … wn. Каждому слову последовательности соответствует множество значений-концептов из онтологической сети – $\{s_{1i}\}\{s_{2i}\} \dots \{s_{ni}\}$. Из каждого множества в процессе контекстного анализа необходимо выбрать по одному значению таким образом, чтобы они находились на максимально близком расстоянии друг от друга. То есть, чтобы сумма расстояний от каждого концепта до всех других была минимальной [5].

Семантическое расстояние между двумя концептами может быть проинтерпретировано как длина кратчайшего пути между соответствующими вершинами в графе онтологической сети. Отдельного рассмотрения заслуживает алгоритм поиска кратчайшего пути в онтологическом графе между узлами концептов слов. Можно ли

строить пути, не учитывая типы связей-отношений между узлами и считая их однотипными? Если нет, то какие последовательности типов связей-отношений в пути можно считать корректными, а какие нет? В зависимости от ответа на эти вопросы можно предложить два подхода к определению семантического расстояния.

1. Простой поиск пути. Тогда решается классическая задача поиска кратчайшего пути в графе. Типы связей-отношений не учитываются. Считается что все дуги одного типа. Еще одним вариантом этого подхода есть числовое ранжирование связей-отношений, где дугам разного типа присваиваются разные весовые коэффициенты, но сам алгоритм поиска кратчайшего пути остается без изменений.
2. Эвристический поиск. При построении кратчайшего пути позволяются только некоторые последовательности типов связей-отношений (например, в цепочке пути разрешена последовательность типов связей *гипернимия-голонимия-гипонимия* и не разрешена последовательность *гипернимия-антонимия-гипонимия*). Такие последовательности предлагается называть эвристиками путей. Процедура поиска кратчайших путей управляется автоматом эвристик, который в качестве фильтра отбирает только те связи-отношения, которые соответствуют заложенным эвристикам.

Когда кратчайший путь найден, его длина принимается в качестве семантического расстояния между данной парой концептов.

Когда на вход блока контекстного семантического анализа поступает пара слов W1 и W2, нужно из множеств их семантических значений S1 и S2 выбрать соответственно пару значений концептов, которой будет соответствовать минимальное семантическое расстояние, то есть минимальная длина кратчайшего пути в сети онтологии. Если построить кратчайший путь в онтологии между лексемами W1 и W2, он пройдет через данную пару концептов, расположенных ближе всего друг к другу. Если на вход поступает последовательность из n лексем, то для каждой из них нужно выполнить (n-1) операцию поиска кратчайшего пути к лексемам-соседям по данному контексту. То есть при решении задачи контекстного анализа входной последовательности длиной в n лексем необходимо выполнить $n(n – 1)/2$ операций поиска. Поиск кратчайшего пути между вершинами в графе является алгоритмически очень сложной операцией, потому данная оценка является, очевидно, крайне неприемлемой.

Нет необходимости строить кратчайшие пути в онтологической сети между всеми лексемами входного предложения. Осуществлять контекстную привязку в онтологии с определением значений концептов лексем нужно, если эти лексемы связаны синтаксическими отношениями в структурах дерева разбора предложения. В случае существования правила, которое связывает некоторую пару слов входной последовательности в единую синтаксическую группу, данные лексемы связываются построением кратчайшего пути между ними в онтологической сети. Среди множеств значений данных лексем выбираются те вершины-концепты, через которые найден кратчайший путь в онтологии. Таким образом, отпадает необходимость построения избыточных цепочек кратчайших путей между всеми словами предложения. Проверяются только те пары лексем, которые имеют связь в синтаксической структуре предложения.

Результаты работы синтаксического анализа учитываются ассоциативно-семантическим контекстным анализом для оптимизации процесса построения ассоциативных связей контекста между словами и словосочетаниями предложения

в иерархической сети онтологической базы знаний. Синтаксическая структура входного предложения является фундаментом и каркасом для этапа семантического анализа.

Но синтаксический анализ, как правило, не в состоянии определить на уровне грамматики однозначную синтаксическую структуру входного предложения. Всегда существует несколько вариантов деревьев вывода входной последовательности предложения, но только один является наиболее адекватным с точки зрения семантической интерпретации синтаксического дерева [6], [7]. Используя описанную технику контекстного семантического связывания концептов-значений в онтологической сети, можно эффективно решать проблемы синтаксической структурной неоднозначности ЕЯ-предложений. Для этого в алгоритмах построения деревьев вывода предложения в местах образования альтернативных связок нужно из набора вариантов выбирать связь, которая объединяет в одну синтаксическую группу наиболее близкие по смыслу концепты-значения. Для этого нужно для каждой синтаксической связи вычислять кратчайшее расстояние в онтологии между концептами-значениями слов, объединяемых данной связью. Значение длины найденного пути принимается за вес построенной связи. Наиболее легкие связи являются более приоритетными на следующем этапе построения дерева.

Таким образом, контекстный семантический анализ управляет процессом синтаксического анализа, а синтаксический анализ генерирует входную структуру для семантического анализа. Мы получаем модель лингвистического процессора, в котором процессы разного уровня анализа являются не столько последовательными, сколько взаимодействующими в параллельно-последовательном режиме, причем управление принадлежит более высокоуровневому процессу – семантическому анализу.

Система текстового мониторинга и качественного оценивания фокусного объекта

Процедуры ассоциативно-семантического контекстного анализа представляют собой функциональное ядро, на основе которого эффективно реализуются различные многоцелевые прикладные системы интеллектуальной обработки естественноязыковых текстов, такие как системы машинного перевода, системы поддержки диалога на естественном языке, программы автоматического реферирования ЕЯ-текстов, тематического индексирования текстов, классификации и рубрикации текстов, кластеризации и фильтрации текстовых потоков и корпусов, ЕЯ-интерфейсы для СУБД и других систем.

Одной из наиболее интересных и одновременно сложных задач, реализованных в рамках данной технологии программирования интеллектуальных ЕЯ-приложений, является построение систем смыслового мониторинга, предназначенных для анализа текстовых потоков и корпусов, с блоком качественного оценивания лингвистических фокусных объектов, выполняющим вычисления качественных характеристик и параметров заданных объектов и процессов. На вход системы подается имя некоторого объекта (или процесса), которое может быть выражено словом, словосочетанием, именем собственным или списком этих слов-имен. Система формирует семантический фокус-образ данного объекта в сети онтологии. После этого производится анализ-сканирование предложений текста с целью поиска качественных оценочных

описаний данного объекта. Для этого сначала выполняется идентификация в тексте всех лексем – упоминаний имени объекта. При этом задействуются функции контекстного ассоциативно-семантического анализа (например, для корректного решения проблемы замены местоимений). Когда определены все вхождения лингвистического объекта в тексте, начинается стадия контекстного поиска элементов качественного оценочного описания в ближайшем окружении мест вхождения. Каждое оценочное понятие онтологии отображается в системе на лингвистическую шкалу типа «хороший – плохой» и получает свое численно-порядковое значение на ней. Отображение «качественных» понятий-концептов в точке лингвистической шкалы и вычисление их фактических значений применительно к объекту происходит динамически в зависимости от локального семантического контекста вхождения объекта.

Важнейшим этапом создания системы семантического мониторинга является формирование лингвистической шкалы для качественных оценочных концептов онтологии. Этот процесс заключается в подборе шкал антонимов (типа «хороший – плохой», «честный – лживый», «умный – глупый» и т.д.), нанесении на эти шкалы промежуточных точек («отлично – хорошо – удовлетворительно – плохо – отвратительно» и т.п.) и соотнесении всех шкал к единой оценочной шкале (условно назовем ее *абсолютная шкала «добро-зло»*), позволяющей системе вычислять интегрированную эмоциональную качественную оценку объекта. В процессе создания лингвистической шкалы было обработано больше трёх тысяч качественных концептов онтологии. Необходимо было нанести их значения на промежуточные антонимические шкалы, а после этого соотнести эти шкалы с нанесенными значениями на абсолютную шкалу. Определение численно-порядковых значений концептов как на промежуточных антонимических шкалах, так и на абсолютной выполняется с помощью ассоциативно-контекстных алгоритмов, которые ищут расстояния в сети онтологии между текущим концептом и концептом максимумом (минимумом) данной шкалы. В зависимости от найденного расстояния определяется точка концепта на шкале и ее численно-порядковое значение. Другой подход, который также был задействован при разработке лингвистической шкалы, использует частотные алгоритмы, определяющие частоту совместного появления пар слов в глобальных корпусах текстов. Принимая гипотезу, что чем чаще пара слов появляется рядом в текстах, тем ближе их семантические значения, частотный анализ был использован для определения значений точек концептов на лингвистической шкале. Необходимо отметить, что слова-антонимы имеют тенденцию часто появляться рядом в текстах, поэтому оценка частоты совместного появления для слов-антонимов будет высокой, что приводит к ощутимому искривлению оценок частотного анализа. Но учитывая этот фактор, можно при обработке корпусов игнорировать значения для слов-антонимов, если в системе есть список пар антонимов для данного языка. Эти два подхода, использованные для построения лингвистической шкалы системы, взаимно компенсировали недостатки друг друга, что сделало разработку шкалы максимально эффективной.

Процесс вычисления качественных оценок лингвистического объекта в тексте не является подсчетом простой суммы абсолютных значений всех качественных оценочных концептов, находящихся в окрестности фокусного объекта. Контекстный ассоциативно-семантический анализ позволяет гибко варьировать значения качественных оценочных концептов в зависимости от локально-глобального контекста.

Это дает возможность учитывать сложные случаи нестандартного (с точки зрения ординарной семантики) применения лексики (например, ироничное «Молодец, Вася! Опять пропустил детский мат!»).

Система позволяет вычислять качественные характеристики и параметры любого заданного лингвистического объекта в корпусах текстов и в текстовых потоках, отслеживая динамику изменений и определяя основные тенденции оценивания фокусного объекта.

Литература

1. Анисимов А.В. Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. – Киев: Наукова думка, 1991.
2. Анисимов А.В., Марченко А.А. Система обработки текстов на естественном языке // Искусственный интеллект. – 2002. – № 4. – С. 157-163.
3. Fillmore Ch.J. The case for case // Universals in linguistic theory / Ed. By E. Bach and B. Halms. – N.Y., 1968.
4. Nirenburg S., Raskin V. Ontological Semantics. – University of New Mexico, 2001.
5. Марченко О.О. Моделювання семантичного контексту при аналізі текстів на природній мові // Вісник Київського університету. – Сер. фіз.-мат. науки. – 2006. – № 3. – С. 230-235.
6. Daniel Jurafsky and James H. Martin (2000). Speech and Language Processing Prentice Hall. – Englewood Cliffs, New Jersey 07632.
7. Ахо А., Ульман Дж. Теория синтаксического анализа, компиляции и перевода. – М: Наука, 1989.

О.О. Марченко, А.О. Никоненко

Контекстний семантичний аналіз тексту. Система текстового моніторингу та якісного оцінювання фокусного об’єкта

Робота описує систему асоціативно-семантичного контекстного аналізу природномовних текстів, на платформі якої реалізовано прикладну систему моніторингу текстових потоків та корпусів з блоком якісного оцінювання лінгвістичних фокусних об’єктів, яка призначена для обчислення різних якісних характеристик та параметрів заданих об’єктів та процесів.

O.O. Marchenko, A.O. Nikonenko

The Contextual Semantic Analysis of Natural Language Text. System of Text Monitoring and Qualitative Estimation of the Focus Object.

This paper describes system of the associative-semantic contextual natural language text analysis. On the base of semantic platform the applied system of monitoring of text streams and corpuses with the block for qualitative estimation of the linguistic focus objects is developed. The block for qualitative estimation calculates various qualitative characteristics and parameters for focus objects and processes.

Статья поступила в редакцию 10.07.2008.